*Nucleic Acids Research, 2004, Vol. 32, Web Server issue* **W253–W256**
*DOI: 10.1093/nar/gkh385*

# CREME: *Cis*-Regulatory Module Explorer for the human genome

**Roded Sharan*, Asa Ben-Hur[1], Gabriela G. Loots[2] and Ivan Ovcharenko[2]**

International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA, [1]Department of Biochemistry, B400 Beckman Center, Stanford University, CA 94305, USA and [2]EEBI and Genome Biology Divisions, L-441 Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA

Received February 12, 2004; Revised and Accepted March 21, 2004

## ABSTRACT

**The binding of transcription factors to specific regulatory sequence elements is a primary mechanism for controlling gene transcription. Eukaryotic genes are often regulated by several transcription factors whose binding sites are tightly clustered and form *cis*-regulatory modules. In this paper, we present a web server, CREME, for identifying and visualizing *cis*-regulatory modules in the promoter regions of a given set of potentially co-regulated genes. CREME relies on a database of putative transcription factor binding sites that have been annotated across the human genome using a library of position weight matrices and evolutionary conservation with the mouse and rat genomes. A search algorithm is applied to this data set to identify combinations of transcription factors whose binding sites tend to co-occur in close proximity in the promoter regions of the input gene set. The identified *cis*-regulatory modules are statistically scored and significant combinations are reported and graphically visualized. Our web server is available at http://creme.dcode.org.**

## INTRODUCTION

Developmental and environmental factors constantly modulate the expression levels of genes in living cells. These changes are primarily triggered by transcription factors (TFs) that physically interact with regulatory sequences located in promoter elements proximal to the transcription start site of a gene. In higher eukaryotes transcriptional regulation is combinatorial in nature: the expression level of a gene is determined by an interplay among several TFs, whose binding sites are organized in a modular fashion along a gene's promoter (1–3). A regulatory element is a sequence segment that contains several spatially clustered transcription factor binding sites (TFBSs), whose corresponding TFs cooperate in the regulation of a group of genes. The set of distinct TFBSs that make up a regulatory element is called a *cis*-regulatory module (CRM). Previous work on identifying modules has followed two main directions. The first has focused on identifying pairs of transcription factors whose binding sites tend to co-occur in the promoter sequences of a group of related genes (4,5). These methods did not require TFBS combinations to occur close together within a region of predefined length. The second direction has focused on identifying regulatory elements of known configuration (6–9). To date, only a few studies have addressed the problem of identifying novel CRMs (10–13).

Recently, we have introduced a new method for identifying novel CRMs and assessing their statistical significance (11). Our method relies on a database of putative TFBSs across the promoters of known human genes which are conserved in mouse and rat. We use evolutionary conservation to increase the reliability of TFBSs predictions: it has been shown that a significant percentage of computationally predicted TFBSs can be reliably eliminated based on evolutionary analysis (14). Proceeding from this database of conserved TFBSs, a search algorithm seeks all combinations of two or more distinct TFBSs that tend to co-occur in a selected set of promoters more frequently than expected by chance. Each CRM is statistically evaluated and significant modules are reported. We have applied this strategy for the analysis of regulation of stress response genes and cell-cycle regulated genes, and have identified several novel CRMs, most of which were shown to be associated with significantly co-expressed or functionally related groups of genes (11).

In this paper, we summarize this strategy for identifying CRMs and describe a web server, CREME (*Cis*-REgulatory Module Explorer), that implements it. The application requires as input a set of putatively co-regulated genes to initiate a search for abundant CRMs in the promoter regions of those genes. A graphical user interface allows users to customize the

*To whom correspondence should be addressed. Tel: +1 510 6662988; Fax: +1 510 6662956; Email: roded@icsi.berkeley.edu
Correspondence may also be addressed to Ivan Ovcharenko. Tel: +1 925 4225035; Fax: +1 925 4222099; Email: ovcharenko1@llnl.gov

search and to visualize the identified CRMs. CREME is available online at http://creme.dcode.org/.

## THE COMPUTATIONAL PIPELINE

The CREME server is designed to identify combinations of TFBSs that tend to co-occur in close proximity in the promoter regions of a specific set of genes. As a preprocessing step we prepare a database of putative TFBSs across the promoter regions of known human genes. Each TFBSs corresponds to a position weight matrix (PWM) match to the human genome that is also conserved in mouse and rat. The module search algorithm proceeds in four phases. First, it identifies non-redundant PWMs whose hits are enriched in the input promoters compared with background set of all other conserved RefSeq promoters. Second, it enumerates all combinations of these PWMs that occur within a window of prescribed length in the input promoters. Third, these combinations are statistically evaluated. Last, significant combinations are reported and visualized.

The database of conserved TFBSs was prepared using PWMs cataloged in the TRANSFAC database (15), which contains over 500 vertebrate TF matrices that comprise about 400 TF families. To reduce the problem of false predictions of binding sites we restricted attention to PWM matches that occur within a sequence segment of 20 bp or more which is highly conserved (>80% sequence identity) in mouse and rat. Genome alignments of human versus mouse and human versus rat were obtained from the ECR Browser, an interactive database consisting of precomputed whole genome alignments from multiple vertebrate genomes (http://ecrbrowser.dcode.org/). These alignments were generated using the most recent assemblies for the three genomes (hg16, mm4 and rn3; http://genome.ucsc.edu/). By overlaying conservation profiles with RefSeq gene annotation mapped to the hg16 human genome assembly, we determined that for the 16 000 non-overlapping RefSeq transcripts, 46% of the promoters (1.5 kb upstream of the transcription start site, or up to the next upstream gene if this region overlaps with another gene) were highly conserved in both the mouse and rat genomes. The rVista 2.0 tool (14) (http://rvista.dcode.org/) was applied to each of the alignments to find TFBSs that are conserved between the two respective genomes. The conservation information from the two rVista computations was superimposed to produce a list of TFBSs that are conserved in the three genomes with PWM similarity scores of 0.8 and above. In total, the CREME database contains 1.4 million putative TFBSs corresponding to 487 different PWMs and spanning 7307 human genes. The addition of rat to the human and mouse alignments reduced the number of computationally predicted TFBS by ~36% while increasing the evolutionary evidence for the functionality of the conserved sites.

Given this database of conserved TFBSs, we aim at identifying all combinations of TFBSs that co-occur in the selected set of promoters. To reduce the large number of possible combinations that need to be considered, we restrict attention to TFBSs that are enriched in the given set of promoters compared with the background set. The enrichment scores are described in (11). We search this filtered data using a hashing technique that identifies all the TFBS combinations that occur in the given promoters. The search is performed using a sliding sequence window of user-defined length. A window is considered to contain a module if it contains at least one binding site for each of the PWMs that make up the module, where the maximum number of PWMs per module is controlled by the user. Abundant CRMs are then statistically scored, based on their number of occurrences, taking into account the frequency of occurrence of their constituent TFBSs across the input promoters, as well as the similarity between the PWM models for the different binding sites.

The significance of a predicted CRM is weighted against the distribution of its occurrences under a null hypothesis that the binding sites for the different PWMs occur independently. This is done via a permutation test, where randomized instances of the input promoters are produced by permuting the identities of the different TFBSs. The frequency of occurrence of a module under the null distribution is approximately normally distributed, where the distribution's parameters are estimated based on these random data sets. Significant combinations are further filtered to eliminate redundant modules, i.e, modules that tend to overlap in their occurrences. The *P*-values of the resulting combinations are adjusted for multiple testing using the *q*-value method (16), which is specifically designed for adjusting the significance of a large number of observations. The reader is referred to (11) for further details on the algorithm and the statistical scoring method.

## USER INTERFACE

CREME requires as input a list of putatively co-regulated or functionally related human genes, provided as either Locus Link (LLID) or GenBank mRNA (NM−) accession numbers (Figure 1, panel 1). The user has control over several search parameters: (i) Hit threshold; (ii) module length; and (iii) number of PWMs per module. The hit threshold parameter controls which putative TFBSs in the database will be included in the analysis. Higher threshold values increase the specificity but decrease the sensitivity of the considered TFBSs. The second parameter specifies the width of the sliding window to be used when searching for modules. Large window sizes increase the sensitivity of the search but may decrease the statistical significance associated with a module. The third parameter determines the maximum number of PWMs that make up a module. The higher the value the more thorough the search at the expense of an increase in processing time.

Upon submitting a query, CREME searches for sets of transcription factors whose binding sites tend to tightly cluster in the promoter regions of the input genes. The identified modules are reported and visualized (Figure 1, panel 2). For each identified module, CREME provides a graphical display that illustrates the promoter regions of the genes in the input set or, alternatively, of only the genes that contain this module. For each promoter, the occurrences of putative binding sites for the PWMs comprising the module are shown, where occurrences of the module are shaded (Figure 2). The display also includes a textual list of these occurrences along with their locations relative to the transcription start site. In addition, the results page contains a link to the list of PWMs whose associated binding sites were found to be enriched in the input promoters.

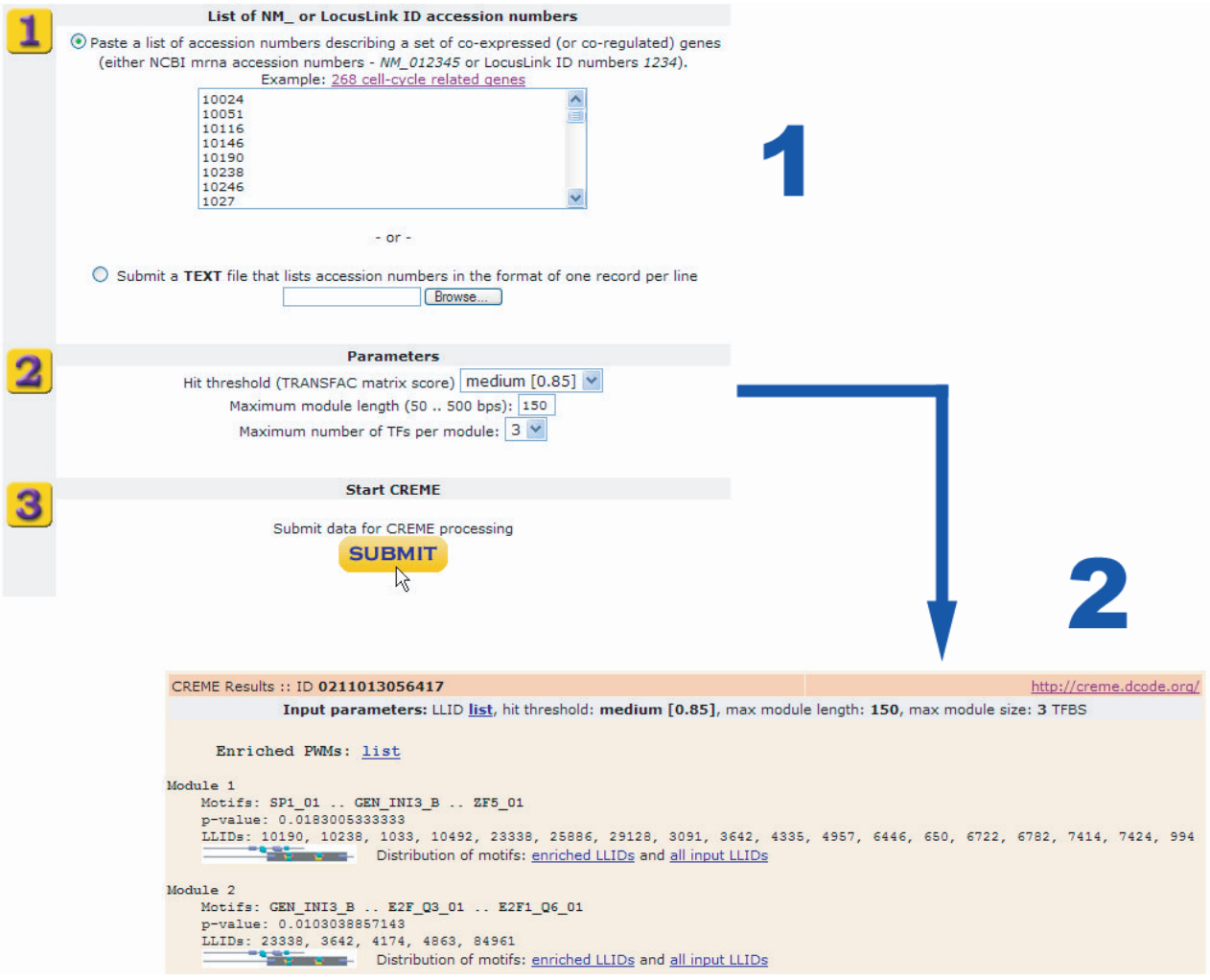


**Figure 1.** The CREME user interface. Upon submission of LLID (or GenBank) accession numbers as pasted text or as a file, and selection of search parameters, the user initiates a CREME process (panel 1) that automatically redirects to the results page (panel 2) when the computation is completed. The results page lists all the detected CRMs whose enrichment in the promoters of the input genes is statistically significant, as well as displaying a link to the list of PWMs that were found to be enriched in the input promoters. For each module, listed are its constituent PWMs, its *P*-value and the LLIDs of genes that contain it. In addition, two visualization links are given for each module, providing a graphical display of the promoters of the input genes and of the subset of genes that contain the module.

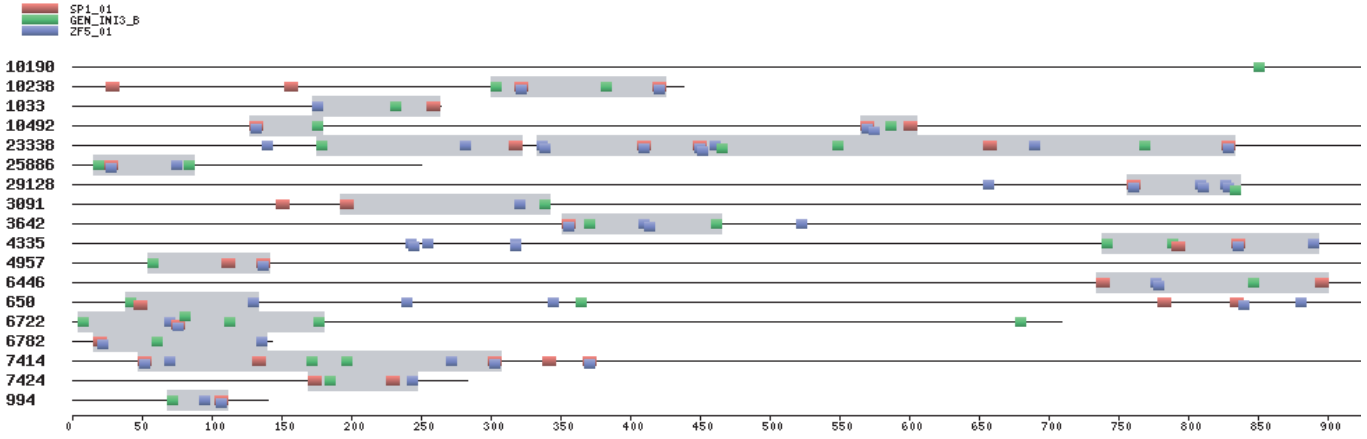


**Figure 2.** CREME visualization of a detected module. Shown are the occurrences of the module and its constituent PWMs in the promoters of the selected set of genes. The module presented in the example consists of three PWMs: SP1_01, GEN_INI3_B and ZF5_01; 18 promoters that contain this module are visualized. SP1_01 sites appear in red, GEN_INI3_B sites in green, and ZF5_01 sites in blue. Module occurrences are shaded in gray. The horizontal scale depicts the distance from each gene's transcription start site (located at the left end). For clarity, only the first 900 bp upstream of the transcription start site are shown.

To illustrate the application of the CREME server, we used it to analyze a set of 268 genes that have been previously shown to be cell-cycle regulated (17) (this set is supplied as a sample input at the website). These genes are a subset of the original set of 651 genes with unique LLID for which conserved promoter segments have been detected. We used the default parameter setting (hit threshold of 0.85; 150 bp-long modules; and at most 3 PWMs per module) for the analysis. The computation resulted in seven significant modules. The first two modules are detailed in Figure 1 (panel 2). The promoter regions of the genes that contain the first module are depicted in Figure 2.

## CONCLUSIONS

We have introduced the CREME web server for identifying *cis*-regulatory modules in the promoters of a given set of genes. The algorithm combines transcription factor binding site model searches, human–mouse–rat evolutionary conservation and statistical assessment of combinations of binding sites. The server reports significantly abundant CRMs along with their *P*-values, and provides a graphical display of these co-occurrences. While there are several other available tools that provide methods for detecting predefined CRMs [see e.g. (7,9)], CREME is able to identify *cis*-regulatory modules *de novo*. Thus, it can assist researchers in the discovery of transcription factors that synergistically activate genes and may, therefore, be responsible for their similar behavior. The identification of such combinatorial modules is critical for understanding how transcriptional regulatory elements are encoded in the human genome, and why certain physiological conditions trigger genes to be turned on or off at the same time.

## ACKNOWLEDGEMENTS

This research was supported in part by NSF ITR Grant CCR-0121555 and Livermore LDRD Grant 04-ERD-052. The work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

## REFERENCES

1. Yuh,C., Bolouri,H. and Davidson,E. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.
2. Ludwig,M., Patel,N. and Kreitman,M. (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development*, **125**, 949–958.
3. Krivan,W. and Wasserman,W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
4. Pilpel,Y., Sudarsanam,P. and Church,G. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153–159.
5. GuhaThakurta,D. and Stormo,G. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
6. Wasserman,W. and Fickett,J. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–81.
7. Frith,M., Spouge,J., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.
8. Sinha,S., vanNimwegen,E. and Siggia,E. D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**(Suppl. 1), I292–I301.
9. Johansson,O., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19**(Suppl. 1), I169–I176.
10. Kel-Margoulis,O., Ivanova,T., Wingender,E. and Kel,A. (2002) Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac. Symp. Biocomput.*, pp. 187–198.
11. Sharan,R., Ovcharenko,I., Ben-Hur,A. and Karp,R. (2003) CREME: a framework for identifying cis-regulatory modules in human–mouse conserved segments. *Bioinformatics*, **19**(Suppl. 1), I283–I291.
12. Segal,E. and Sharan,R. (2004) A discriminative model for identifying spatial cis-regulatory modules. In Gusfield,D., Bourne,P., Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, ACM Press, San Diego, CA, pp. 141–149.
13. Marsan,L. and Sagot,M.-F. (2000) Extracting structured motifs using a suffix tree—algorithms and application to promoter consensus identification. In Shamir,R., Miyamo,S., Istrail,S., Pevzner,P. and Waterman,M. (eds) *Proceedings of the Fourth Annual International Conference on Research in Computational Molecular Biology*, ACM Press, New York, NY, pp. 210–219.
14. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin, E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.
15. Wingender,E., Chen,X., Hehl,R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I. and Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
16. Storey,J. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
17. Whitfield,M., Sherlock,G., Saldanha,A.J, Murray,J., Murray J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.